УДК 543.42

ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА СПЕКТРОВ В ЛАЗЕРНО-ФЛУОРЕСЦЕНТНОМ ИССЛЕДОВАНИИ СМЕСЕЙ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ[1]

В.Б.Борисов, В.М.Немец, М.Н.Полянский, А.А.Соловьев
НИИ физики Санкт-Петербургского государственного университета
198904, Санкт-Петербург, Ульяновская, 1

Исследованы возможности применения метода распознавания образов (метода главных компонент) для решения задачи идентификации сложных органических соединений по спектрам их лазерной флуоресценции. Предложена методика двухстадийного применения метода главных компонент для анализа сложных соединений. Показана его работоспособность на примере идентификации бензинов с различными октановыми числами.

Немец Валерий Михайлович – доктор технических наук, профессор, заведующий лабораторией спектрального анализа НИИФизики Санкт-Петербургского государственного университета.
Область научных интересов: спектральный анализ газов, жидких и твердых веществ и материалов.
Автор 230 опубликованных работ.

Борисов Валерий Борисович – кандидат физико-математических наук, старший научный сотрудник НИИФизики Санкт-Петербургского государственного университета.
Область научных интересов: физика плазмы, лазерный спектральный анализ газов.
Автор 58 опубликованных работ.

Соловьёв Анатолий Анатольевич – кандидат технических наук, старший научный сотрудник НИИФизики Санкт-Петербургского государственного университета.
Область научных интересов: спектральный анализ, лазерная аналитика.
Автор 160 опубликованных работ.

Полянский Михаил Николаевич – аспирант физического факультета Санкт-Петербургского государственного университета.
Область научных интересов: лазерный спектральный анализ жидких сред.
Автор 5 опубликованных работ.

Введение

Общеизвестна роль оптических спектральных методов в современной аналитике. Это, прежде всего, методы, базирующиеся на атомной и молекулярной спектроскопии, решающие значительное число задач анализа самых различных веществ и материалов.

Ограничения в развитии оптических спектральных методов в последнее время связаны в основном со сложностью или невозможностью непосредственной интерпретации спектров с большим (больше двух) числом перекрывающихся линий или полос. Классическим вариантом преодоления такого рода трудностей всегда являлось применение различных методов пробообработки, что существенно усложняло и удорожало анализ, делало его лабораторным, т.е. неоперативным, что исключает, например, возможность использования метода для контроля технологических процессов.

В связи с появлением и распространением компьютеров последних поколений появилась во многих случаях возможность снятия упомянутых выше ограничений на основе разработки методов компьютерного анализа сложных спектральных кривых. Сегодня такие методы могут быть

[1] Работа выполнена при поддержке программы "Университеты России".

условно разделены на две группы. Первая группа – методы выделения отдельных полос из суммарного спектра и определения соответствия компонентов спектра компонентам исследуемого вещества. Методы первой группы хорошо работают в случае наличия априорной информации о форме спектров составляющих данное соединение компонентов и небольшого количества таковых (классический случай – две перекрывающиеся гауссовы кривые). Ко второй группе относятся методы анализа спектральных кривых как таковых, вне связи с природой спектра. В данном случае оценивается степень "похожести" спектра исследуемого образца на спектры эталонных образцов. Использование методов второй группы предполагает проведение статистического анализа большого набора спектральных кривых веществ с известным составом и выделение на этой основе некоторого ограниченного набора параметров, по которым будет производиться сравнение. Следует заметить, что данные методы по существу являются методами "распознавания образов", активно разрабатываемыми в настоящее время в применении к таким задачам, как распознавание изображений (рукописного текста, фотопортретов, отпечатков пальцев) и распознавание речи. Однако в спектральном анализе эти методы пока не получили широкого распространения, несмотря на интересные результаты, полученные по крайней мере в одной известной нам работе [1]. По всей видимости, это обусловлено некоторой консервативностью в подходе к спектроскопии как к инструменту точного непосредственного определения состава веществ по спектрам отдельных компонентов.

Постановка задачи

В данной работе поставлена задача исследования возможностей метода главных компонент, являющегося базовым в задачах распознавания образов в применении к спектрам лазерной флуоресценции органических соединений. В качестве объектов исследований были выбраны бензины и машинные масла. Выбор обусловлен следующими причинами. Во-первых, в настоящее время существует насущная необходимость в оперативных методиках контроля нефтеперегонных процессов и масляных загрязнений почвы и водоемов. Во-вторых, флуоресцентный метод позволяет производить дистанционное зондирование, что немаловажно для указанных задач. И, наконец, использование бензинов очень удобно для тестирования предлагаемого метода обработки спектров, поскольку существует параметр (октановое число), по которому может быть произведена классификация исследуемых образцов.

Разумеется, выбранный объект исследования (спектры флуоресценции бензинов) не является единственно возможным применением метода главных компонент в спектральном анализе. Разрабатываемые на данном примере алгоритмы могут быть в дальнейшем применены для обработки большого класса сложных спектров различной природы, трудно поддающихся непосредственной интерпретации.

Экспериментальные исследования

Для выполнения работы была сконструирована экспериментальная установка (рис. 1). Флуоресценция в исследуемом веществе, находящемся в кварцевой кювете 2, возбуждается импульсным азотным лазером 1 (длина волны генерации 337 нм, мощность в импульсе 20 кВт, длительность импульса 3 нс). Сигнал флуоресценции, прошедший через монохроматор 3, регистрируется фотоумножителем 4, усиливается и усредняется по 100 – 1000 импульсам возбуждения для уменьшения влияния нестабильности лазера.

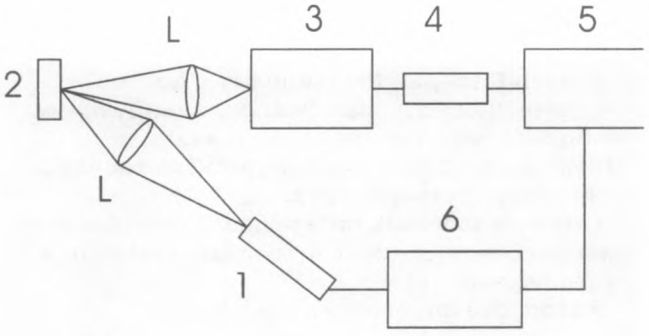

Рис.1. Схема экспериментальной установки:
1 – лазер; 2 – кварцевая кювета с исследуемым веществом; 3 – монохроматор; 4 – фотоумножитель; 5 – схема регистрации; 6 – схема управления лазером; L – фокусирующие объективы

Уже предварительные эксперименты, проводившиеся с меньшим спектральным разрешением без накопления информации за большое количество импульсов, показали, что в спектрах флуоресценции бензинов и масел наблюдаются определенные различия [2]. В данной работе были проведены по возможности более точные измерения спектров флуоресценции, что и позволило провести их обработку по методике, описанной ниже.

Были получены спектры флуоресценции нескольких машинных масел и набора бензинов. Образцы представлены на рисунке (рис. 2) (спектры нормированы по максимальному значению сигнала).

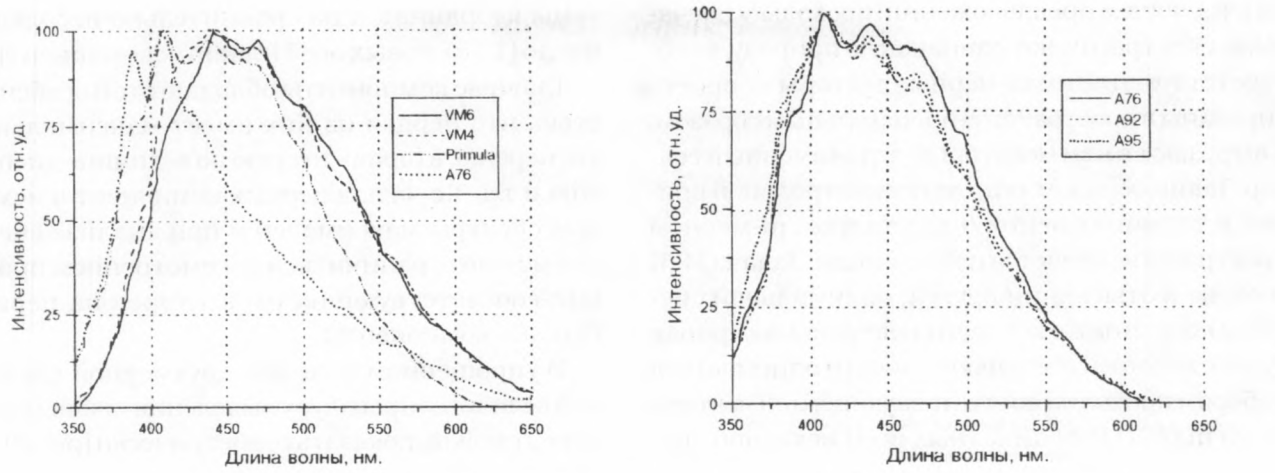

Рис.2. Спектры флуоресценции машинных масел и бензинов

Сопоставление спектральных кривых показывает, что существуют определенные различия в спектрах флуоресценции близких по составу соединений, однако сложная форма кривых не позволяет производить непосредственную идентификацию исследуемых образцов по каким-либо характерным особенностям спектра. В дальнейшем производился анализ полученных спектров с использованием метода главных компонент. Остановимся на этом подробнее.

Методика обработки данных

Идея обсуждаемой методики обработки данных заключается в отыскании такой схемы преобразования исходных спектров, которая позволит выделить некие обобщенные особенности, по которым в дальнейшем может быть произведена классификация исследуемых веществ.

Метод главных компонент (МГК, компонентный анализ) является одним из методов многомерного статистического анализа (МСА). МСА обладает широкими возможностями в отображении и моделировании реальных явлений и процессов, изначально имеющих многопризнаковую природу. Все новейшие разработки, посвященные проблемам приложения нечетких множеств, моделирования катастроф, распознавания образов, сценарного прогнозирования и т.д., предполагают многомерное представление наблюдаемых объектов.

МСА следует рассматривать как логическое развитие методов традиционной статистики. Принципиальное различие заключается в том, что объекты рассматриваются здесь с учетом не одного-двух, а одновременно некоторого множества признаков. Это позволяет добиваться в исследованиях полноты описания наблюдаемых объектов и объективности последующих выводов.

Понятие МСА может быть сформулировано следующим образом: "*Это совокупность глубоко формализованных статистических методов, базирующихся на представлении исходной информации в многомерном геометрическом пространстве и позволяющих определять неявные (латентные), но объективно существующие закономерности в структуре и тенденциях развития изучаемых явлений и процессов*" [3]. Практическое применение методов МСА требует обязательного использования вычислительной техники. Можно сказать, что эти методы в силу сложности и трудоемкости не реализуемы без привлечения технических средств. Широкое распространение МСА в исследованиях началось именно с появлением первых ЭВМ.

Из числа методов, позволяющих обобщать значения элементарных признаков, метод главных компонент выделяется простой логической конструкцией и сравнительно высокой наглядностью. МГК дает возможность по m – числу исходных признаков выделить m главных компонент или обобщенных признаков. Пространство главных компонент ортогонально.

Поскольку в данной работе МГК применяется для анализа спектральных кривых, дальнейшее изложение ведется на примере данной конкретной задачи.

В терминах многомерной статистики любая спектральная кривая, описываемая фактически дискретным набором значений амплитуд оптического сигнала на некоторой последовательности длин волн, может быть представлена вектором в m-мерном пространстве, где m – число дискретных измерений спектра. В нашей работе для анализа спектров флуоресценции использовался отрезок спектральной кривой 350 – 550 нм с шагом 1 нм, т.е. спектральная кривая описывалась набором из 201 характеристики, или вектором в 201-мерном пространстве. МГК базируется

на следующем предположении: поскольку изучаемые спектры имеют одинаковую природу, то соответствующие им m-мерные векторы не ориентированы в пространстве произвольным образом, а вырезают в нем некоторый ограниченный сектор. Таким образом, описание спектральной кривой в терминах вектора в исходном m-мерном пространстве является избыточным. Задача МГК состоит в отыскании другой, оптимальной системы координат, в которой спектральная кривая будет с хорошей степенью точности описываться набором проекций соответствующего ей вектора не на m (201) координатных осей исходной системы координат, а на сравнительно небольшое число (1 – 3) новых осей (главных компонент).

Главные компоненты обладают той особенностью, что первая из них имеет максимальную дисперсию, вторая – вторую по величине дисперсию и т.д. Т.е. вклад первых компонент в исходные спектры максимален и при анализе спектров можно ограничиться рассмотрением проекций соответствующих им векторов на первые главные компоненты.

В упрощенном виде для двухмерной случайной величины процедуру выделения главных компонент можно показать геометрически (рис.3).

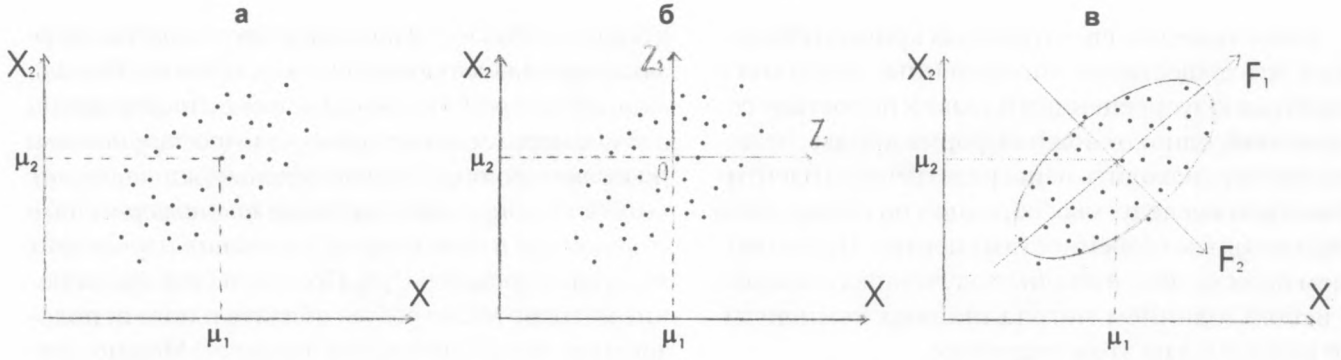

Рис.3. Геометрическая интерпретация процедуры выделения главных компонент:
а-первоначально имеется некоторое эмпирическое распределение данных в двумерном признаковом пространстве X_1, X_2 с центром (μ_1; μ_2); б-система координат переносится в центр распределения данных. Новые координаты Z_1, Z_2; в-находятся параметры эллипса, описывающего эмпирическое распределение данных, соответственно устанавливается положение главных компонент F_1, F_2

Математически процедура выделения главных компонент заключается в следующем:

а) исходный набор спектральных кривых представляется в форме матрицы X с размерностью n x m; n – число спектров, m – число дискретных измерений каждого спектра;

б) перенос начала системы координат в центр распределения данных осуществляется вычитанием усредненной спектральной кривой из каждого спектра:

$$z_{ij} = x_{ij} - < x_j > ;$$

в) вычисляется ковариационная матрица S:

$$S = \frac{1}{n} Z^T Z ;$$

г) нормированные собственные векторы матрицы S и являются ортами искомой системы координат (главными компонентами), причем большей дисперсией обладают главные компоненты, соответствующие большим собственным значениям.

После этого находятся проекции исходных спектров на оси новой системы координат.

В данной работе была предпринята попытка многостадийного применения метода главных компонент. Идея многостадийного анализа заключается в следующем. На первом этапе обработки спектральных сигналов в анализе участвуют все имеющиеся спектральные кривые; здесь происходит приблизительное разделение исследуемых веществ на группы, имеющие схожие формы спектров. Поскольку на выбор новой системы координат на этом этапе оказывает влияние набор спектральных кривых, имеющих сравнительно широкий разброс в форме и характерных особенностях, то, очевидно, при описании спектров первыми главными компонентами, может происходить значительная потеря информации, заключенной в отбрасываемых главных компонентах. На рисунке (рис. 4) представлен пример восстановления исходного спектра бензина А-76 по двум первым главным компонентам ГК1 и ГК2 на первом этапе. Видно, что полнота описания в значительной степени потеряна. Первый этап многостадийного метода главных компонент является, по сути, обычным методом главных компонент.

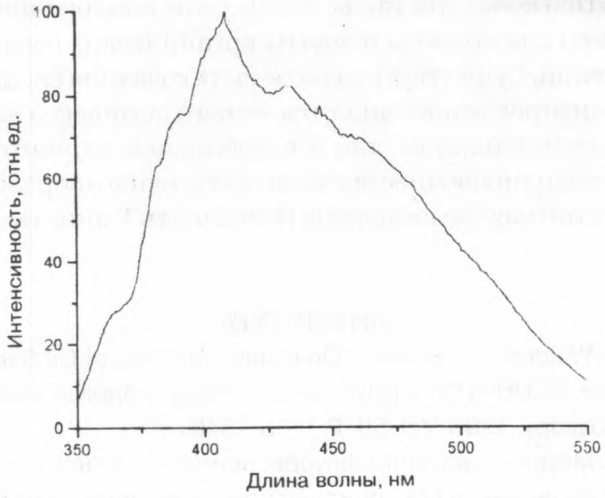

Рис.4. Исходная (сплошная) и восстановленная (пунктир) по двум первым главным компонентам, выделенным на первом этапе, спектральная кривая

По результатам первого этапа производится выделение из исходного набора спектров групп спектров, занимающих ограниченные области в пространстве первых главных компонент, выделенных на первом этапе. Эти спектральные кривые имеют менее широкий разброс форм, и соответственно их векторы занимают в исходном n-мерном пространстве более узкий сектор, чем весь исходный набор спектров. Теперь метод главных компонент применяется отдельно для каждой такой группы. Очевидно, что с уменьшением разброса форм исходных спектров, участвующих в выборе новой системы координат, происходит увеличение точности описания спектров первыми главными компонентами. На следующем рисунке (рис. 5) представлена восстановленная по двум первым главным компонентам, выделенным на втором этапе, спектральная кривая того же бензина.

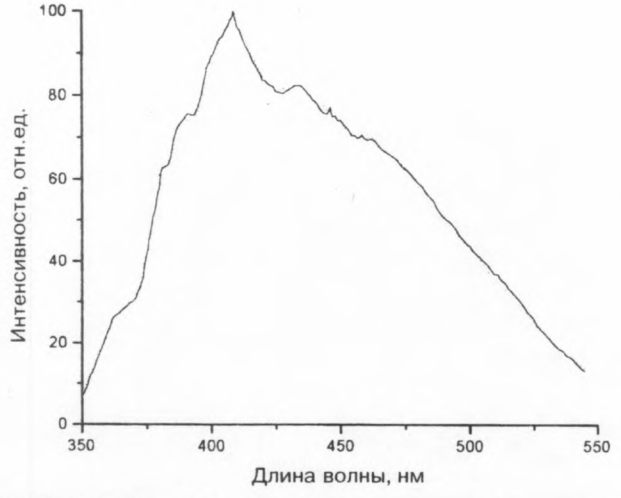

Рис.5. Исходная (сплошная) и восстановленная (пунктир) по двум первым главным компонентам, выделенным на втором этапе, спектральная кривая

Среднеквадратичное отклонение восстановленной спектральной кривой от исходной на первом этапе составляет 25.94, на втором – 15.15. Таким образом, на втором этапе становится возможным проводить более детальную классификацию исследуемых образцов.

В данной работе была применена двухстадийная схема обработки данных, однако, очевидно, при решении некоторых сложных задач может оказаться целесообразным применение трех и более стадий для более точного описания исследуемых процессов.

Результаты

Для использования метода главных компонент было разработано специальное программное обеспечение. Анализ производился в два этапа. На первом этапе рассматривался весь имеющийся набор спектров флуоресценции бензинов. Получена следующая картина распределения данных в пространстве главных компонент (рис. 6).

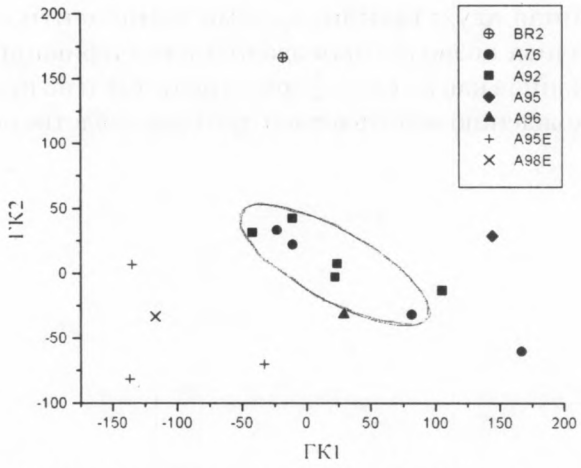

Рис.6. Результаты первой стадии обработки спектров методом главных компонент

Видно, что таким образом могут быть однозначно выделены "Е" – бензины (бензины различных европейских производителей). Стоит заметить, что именно эти бензины не удается определять рефрактометрическим методом, также разрабатываемым в нашей лаборатории [4]. Очень заметно выделяется бензин БР-2 ("галоша") с октановым числом 70, получаемый прямой перегонкой нефти без добавления присадок. Относительно российских бензинов можно сказать, что, по всей видимости, различия обусловлены по большей части различиями сортов нефти, используемой для производства бензинов, а не октановым числом. Поэтому на втором этапе анализ производился только с использованием нескольких бензинов, занимающих на рис.6 огра-

ниченную область (на рисунке очерчено замкнутой линией). Сюда входят четыре 92-х и три 76-х бензина. Оказалось, что в этом случае бензины с разными октановыми числами образуют компактные группы (рис. 7).

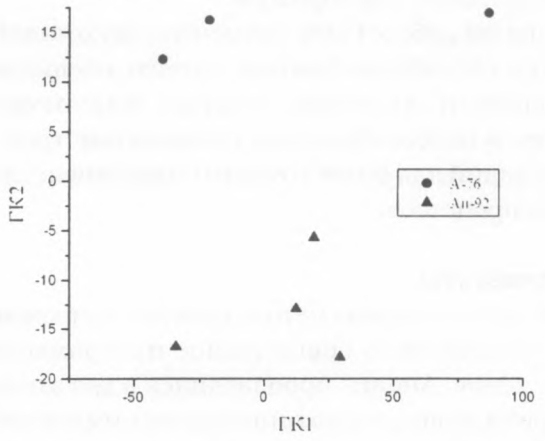

Рис.7. Результаты второй стадии обработки спектров методом главных компонент

Таким образом, по всей видимости, предложенная двухстадийная схема компонентного анализа позволит производить классификацию бензинов как по октановому числу, так и по происхождению используемой при производстве нефти. Возможно также построение аналогичных схем для анализа и других органических соединений. Существует возможность создания специализированного аналитического комплекса как в стационарном, так и в мобильном варианте. Ориентировочное время анализа одного образца в этом случае не должно превышать 1 минуты.

ЛИТЕРАТУРА:

1. Waldemar I. Friesen Qualitative Analysis of Oil Sand Slurries Using On-line NIR Spectroscopy // Applied Spectroscopy. 1996. Vol. 50. P.1535-1540.

2. Методика лазерно-флуоресцентного анализа нефтепродуктов и обнаружения примесей органических соединений в воде/В.Б.Борисов, В.М.Немец, А.А.Пастор, М.Н.Полянский, А.А.Соловьев // Приборы и системы управления. 1999. № 6. С.43-44.

3. Многомерный статистический анализ в экономике/Л.А.Сошникова, В.Н.Тамашевич, Г.Уебе, М.Шефер. М.: Юнити, 1999. 598 с.

4. Рефрактометрический определитель марки бензинов/В.В.Берцев, В.Б.Борисов, Л.В.Гатилова, В.М.Немец, Ю.А.Пиотровский, А.А.Соловьев // XIV Уральская конференции по спектроскопии: Тезисы докладов. Заречный, 1999. С.146.

* * * * *